Learning to Find Usability Problems in Internet Time

Mikael B. Skov and Jan Stage
Department of Computer Science
Aalborg University
Aalborg Øst, Denmark
+45 9635 8080
{dubois, jans}@cs.auc.dk

ABSTRACT
This article reports from an empirical study of web site usability testing conducted by a diverse group of people with no formal education in software development or usability testing. 36 teams of four to eight first-year university students with interest but no education in information technology were trained in a simple approach to web-site usability testing that can be taught in less than one week. The article describes how they applied this approach for planning, conducting, and interpreting a usability evaluation of the same web site. It is concluded that basic usability testing skills can be developed in a week. The student teams gained competence in defining good task assignments and ability to express the problems they found. They were less successful when it came to interpretation and analytical skills. They found quite few problems, and they seemed to lack an understanding of the characteristics that makes a problem list applicable.

Keywords
Usability test, web-site development, usability engineering

INTRODUCTION
The fusion of emerging information technologies will facilitate new patterns of work and interaction and thereby impose a variety of radical changes to our daily life. These changes will create new demands that have been summarized in the requirement for information technologies to be available and accessible to anyone, anywhere, anytime. The World Wide Web is a significant contribution to fulfill this requirement. Today, the Web makes substantial amounts of information available to anyone with a computer that is connected to the Internet. We are still far from the ideal, but the Web has definitely brought us a step forward.

The Web has, however, also brought us a step backward. The work practices that are presently employed in web-site development seem to largely ignore the body of knowledge and methodologies that has been established over the last decades in the disciplines of software engineering, human-computer interaction, and usability engineering [16].

Time pressure is a key reason why established knowledge and methodologies are ignored [2]. Web developers experience a strong push for speed. Users of web sites rapidly change preferences and patterns of use, and new ideas for design and functionality emerge constantly. This makes customers and management demand development cycles that are considerably shorter than in traditional software development [1, 4].

Lack of knowledge is another important reason. Many web sites are designed and implemented by multidisciplinary teams that involve such diverse professions as information architects, web developers, graphic designers, brand and content strategists, etc. Such teams are usually not familiar with established knowledge on human-computer interaction [3].

This lack of knowledge excludes usability issues from web-site development. Development teams focus on site and page design, functionality, and technical implementation, whereas usability work is largely ignored [2]. In addition, the strong limitations in terms of development time effectively prohibits usability testing in the classical sense, where it is conducted systematically by experienced test teams in sophisticated laboratories, as it is described in [5, 11, 13, 14].

There are several suggestions aiming to overcome the challenges and time pressure that are typical for web-site development, e.g. [1, 3, 4]. Yet the focus in all of these proposals is on analysis and design. There is not a similar stream of techniques to support integration of usability issues in development and processes for enhancing practical skills with usability testing.

This article reports from an empirical study of the potential for dissemination of fundamental usability engineering skills. We have tried to provide people who have an interest in information technology but no formal education in software development or usability engineering with a simple approach to usability testing, and we have tried to do this in less than a week. Below, we start by describing heuristic inspection which was an earlier attempt to reduce the effort needed to conduct usability testing. Our approach is presented as a complementary attempt to reduce usability testing efforts, and we provide results from an empirical
study where 234 first-year university students learned and used the approach to conduct a usability test.

HEURISTIC INSPECTION
A main problem in planning and conducting full-scale usability tests is the integration of users. Considerable costs arise when a large group of users is involved in a series of tests. The approach denoted as heuristic inspection evolved as an attempt to reduce these costs [9, 10, 12].

The basic idea in heuristic inspection is that an interface design is evaluated by relating it to a set of guidelines, called heuristics [12]. The first heuristics consisted of nine principles [9], that have been developed further over the last ten years. The aim of the heuristics is to equip people who are not usability specialists to conduct heuristic inspections. Some of the empirical studies of the approach have been based on university students or readers of a computer magazine who act as evaluators [12].

The literature on heuristic inspection also includes empirical studies of its capability for finding usability problems. The first studies were conducted by the authors of the approach. Based on these studies it was concluded, that 5 inspectors were able to find between 55 and 90 percent of the usability problems in four different interfaces [9], and major problems have a higher probability of being found than minor [10].

Other studies have produced less promising results. The classical usability test yields similar or better results compared to inspections conducted by individuals [7]. Moreover, heuristic inspection tends to find many low-priority problems [6]. The best results with heuristic inspection seem to be obtained when conducted by user interface specialists [7, 10].

The idea behind heuristic inspection is to achieve a simplified way of conducting usability tests. However, the empirical results indicate that we move the problem from finding users to finding user interface specialists. For a small organization developing web-based systems both of these problems may be equally hard to overcome.

On a more general level the relevance of heuristic inspection can also be questioned. It has been argued that real users is an indispensable prerequisite for usability testing. If they are removed, it is at the expense of realism [14].

The fundamental idea of heuristic inspection is to reduce testing effort by making developers or other kinds of non-users invent problems that users might experience. This idea has to some extent inspired our approach to usability testing that is presented in the following section.

USABILITY TESTING APPROACH
The approach to usability testing was developed through a course that was part of a curriculum for the first year at Aalborg University, Denmark. The overall purpose of the course was to teach and train students in fundamentals of computerized systems with a particular emphasis on usability issues. The course included ten class meetings as illustrated in figure 1, each lasting four hours that was divided between two hours of class lectures and two hours of exercises in smaller teams. All class meetings, except number one and five, addressed aspects of usability and web sites. The course required no specific skills within information technology that explains the introduction of course number one and five. The purpose of the exercises was to practice selected techniques from the lectures. In the first four class meetings, the exercises made the students conduct small usability pilot tests in order to train and practice their practical skills. The last six exercises were devoted to conducting a more realistic usability test of a web site.

A number of techniques for usability testing were presented in the course. The first of two primary techniques was the think-aloud protocol, which is a technique where test subjects are encouraged to think aloud while solving a set of tasks by means of the system that is tested, cf. [11]. The second technique is based on questionnaires that test subjects fill in after completing each task and after completion of the entire test, cf. [15]. Additional techniques such as interviewing, heuristic inspection, cognitive walkthroughs, etc. were additionally briefly presented to the students.

The tangible product of a usability evaluation is a usability report, that identifies usability problems of the product, system, or web site in question. It was suggested that a usability report should consist of an executive summary (1 page), description of the approach applied (2 pages), results of the evaluation (5-6 pages), and a discussion of methodology (1 page). It was also emphasized that the problems identified should be categorized, at least in terms of major and minor usability problems. In addition, a report should include all data material collected such as log-files, tasks for test subjects, questionnaires etc. A prototypical example of a usability report was given to the students for inspiration.

EMPIRICAL STUDY
We have made an empirical study of the usability approach that was taught to the students.

Participants
This study involved 36 teams of first year university students who used the approach to conduct a usability evaluation of the email services at the Hotmail web-site. The 36 teams consisted of 234 students in total, of which 129 acted as test subjects, from such diverse educations as architecture and design, informatics, planning and environment and chartered surveyor. These studies are all part of a natural science or engineering program at Aalborg University. Figure 7 at the back of the paper describes the 36 teams that participated in the study.
Data Collection
Each student team was required to apply at least one of the two primary techniques, and they were allowed to supplement this with other techniques according to their own choice. With both techniques, the team should among themselves choose a test monitor and a number of loggers (they were recommended to include two loggers), who should examine the system, design task assignments for the test subjects, and prepare the test, cf. [14]. The rest of each team acted as test subjects, and the web site used for testing was kept secret to them.

Each team was given a very specific two-page scenario stating that they should conduct a usability test of the Hotmail web-site (www.hotmail.com). The scenario included a list of features that emphasized the parts of Hotmail they were supposed to test. Each usability test session was planned to last approximately one hour. Due to the pedagogical approach of the university, each team has its own office. Most teams conducted the tests in this office, which was equipped with a personal computer and Internet access. After the test, the entire team worked together on the analysis and identification of usability problems and produced the usability report.

Data Analysis
The usability reports were the primary source of data for our empirical study. All reports were analyzed, evaluated, and marked by both authors of this paper. First, we worked individually and marked each report in terms of 17 different factors. The markings were made on a scale of 1 to 5, with 5 being the best. Second, all reports and evaluations were compared and a final evaluation on each factor was negotiated. In this article, we focus on the following five factors: #1 the planning and conduction of the evaluation, #2 the quality of the task assignments, #3 the clarity and quality of the problems listed in the report, #4 the practical relevance of these problems, and #5 the number and relevance of the usability problems identified. In order to provide a basis for comparison, we have also employed usability reports produced by teams from eight professional laboratories, cf. [8]. These teams have evaluated the same web site according to the scenario mentioned above, and their reports were analyzed, evaluated, and marked through the same procedure as the student reports.

Validity
The specific conditions of this study limit its validity in a number of ways. First, the environment in which the tests were conducted was in many cases not optimal for a usability test session. In some cases, the students were faced with slow Internet access that influenced the results. Second, motivation and stress factors could prove important in this study. None of the teams volunteered for the course (and the study) and none of them received any

<table>
<thead>
<tr>
<th>No.</th>
<th>Topic and Content of Class Lecture</th>
<th>Exercise</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Introduction and computer networks</td>
<td>Pilot test: The student teams conducted simple pilot usability tests of web sites to train their practical skills within usability test conduction. The student teams chose the tested web site themselves and experiences with the conduction and the results were discussed afterwards with the course organizer.</td>
</tr>
<tr>
<td>2</td>
<td>Usability issues: guidelines and measurement: Basic introduction to usability issues and guidelines for interaction design</td>
<td>Pilot test:</td>
</tr>
<tr>
<td>3</td>
<td>Usability testing: think-aloud protocol: The think-aloud protocol and how to set up a test scenario. User groups and their different needs.</td>
<td>Pilot test:</td>
</tr>
<tr>
<td>4</td>
<td>Usability testing: questionnaires: How to apply questionnaires as a mean for collecting data and how to use different kinds of questions</td>
<td>Pilot test:</td>
</tr>
<tr>
<td>5</td>
<td>Computer architecture</td>
<td>Pilot test:</td>
</tr>
<tr>
<td>6</td>
<td>Usability testing: data analysis: Hot to describe the usability testing method and how to collect and analyze the empirical data.</td>
<td>Pilot test:</td>
</tr>
<tr>
<td>7</td>
<td>Usability issues: techniques and documentation: Other usability evaluation techniques and how to conduct a full size usability test session in a laboratory</td>
<td>Pilot test:</td>
</tr>
<tr>
<td>8</td>
<td>Web-site: usability: Different kinds of web structures and similarities and differences between information search and web surfing</td>
<td>Pilot test:</td>
</tr>
<tr>
<td>9</td>
<td>Web-site: orientation and navigation: Basic guidelines for web design and principles for orientation and navigation</td>
<td>Pilot test:</td>
</tr>
<tr>
<td>10</td>
<td>Web-site: visual design: Principles for visual design and different interaction styles</td>
<td>Pilot test:</td>
</tr>
</tbody>
</table>

Figure 1. The usability testing approach presented in 10 lectures and the corresponding exercises
payment or other kind of compensation; all teams participated in the course because it was a mandatory part of their curriculum. This implies that students did not have the same kinds of incentives for conducting the usability test sessions as people in a professional usability laboratory. Finally, the demographics of the test subjects are not varied with respect to age and education. Most test subjects a female or a male of approximately 21 years of age with approximately the same school background and recently started on a design-oriented education. The main difference is the different curricula they follow.

USABILITY TEST RESULTS

We have identified five factors that illustrate similarities and differences in the approaches to the usability testing sessions and the results of the testing sessions between the student teams and the professional laboratories.

A key aim in usability testing is to uncover and identify usability problems, and this is the third factor in this study with differences, cf. figure 4. The student teams are on average able to find 7.9 problems. They find between 1 and 19 problems with half of the teams finding between 6 and 10 problems. Thus the distribution seems to be reasonable, but compared to the professionals there is a clear difference. The average for the professionals is 23.0 problems identified; only one of them scores in the same group as a considerable number of student teams; that is between 8 and 12 problems. Only one student team identified a number of problems that is comparable to the professional laboratories.

Figure 3. Conduction of the usability test.

The first factor is the task factor where the distribution of markings is illustrated in figure 2. This factor measures the relevance of the tasks, the number of tasks, and the extent to which they cover the areas specified in the scenario. The student teams cover all five elements of the scale with a few at the top and at the bottom. The average is 3.3 for all student teams, which is a little above the middle. The professional laboratories score almost the same result, with distribution across the whole scale and an average of 3.5. This is by no means impressive for the professionals; thus the comparable result produced by the students is rather due to a general low quality of the tasks defined by the professionals.

The second factor is the conducting of tests, which reflects how well the tests were planned, organized, and carried out, cf. figure 3. The majority of student teams score 4, which indicates well-conducted tests with a couple of problematic characteristics. The average on 3.7 also reflects the general quality of the test processes. The professional laboratories score an average of 4.6 on this factor, and 6 out of 8 score the top mark. This is as it should be expected because experience will tend to raise this factor. However, the student teams perform rather well with respect to planning and conducting the usability testing sessions.

Figure 2. Number and quality of designed tasks.

A different factor that exhibits a difference is the practical relevance of the problem list, cf. figure 5. The student teams are almost evenly distributed on the five marks of the scale, and their average is 3.2. Yet when we compare these to the professional laboratories, there is a clear difference. The professionals score an average of 4.6 where 6 out of 8 laboratories score the top mark. This difference can partly be explained from the experience of the professionals in expressing problems in a way that make them relevant to their customers. Another source may be that the course has focused too little on discussing the nature of a problem; it has not been treated specifically with examples of relevant and irrelevant problems.

Figure 4. Number of identified usability problems
A second factor with a broad distribution is the clarity of the problem list, cf. figure 6. This factor measures how well each problem is described, explained, and illustrated and how easy it is to gain an overview of the complete list of problems. Here, the student teams are distributed mainly around the middle of the scale with a few at the top and at the bottom. Their average is 2.9, just below the middle. Compared to the professional laboratories, they are also doing quite good. The professionals are distributed from 2 to 5 with an average of 3.5. Again, this is not impressive for professional laboratories.

Figure 5. Practical relevance of identified problems

Most of the teams found too few problems. It was also difficult for them to express the problems found in a manner that would be relevant to a practicing software developer.

The idea of this approach is to reduce the efforts needed to conduct usability testing. This is consistent with the ideas behind heuristic inspection and other walkthrough techniques. On a more general level, it would be interesting to identify other potential areas for reducing effort.

The simple approach to usability testing did provide the students with fundamental skills in usability engineering. Thus it is possible to have usability work conducted by people with primary occupations and competencies that are far away from software development and usability engineering. We see the approach as a valuable contribution to the necessary development emphasized here: “Organizations and individuals stuck in the hierarchies and rigidity of the past will not foster what it takes to be successful in the age of creativity, the age of the user, and the age of the Internet economy” [1].

ACKNOWLEDGMENTS

CONCLUSION

The existing low level of skills in usability engineering among web-site development teams is likely to prohibit moves towards the ideal of anyone, anywhere, anytime. This article has described a simple approach to usability testing that aims at quickly teaching fundamental usability skills to people without any formal education in software development and usability engineering. Whether this approach is practical has been explored through a large empirical study where 36 student teams have learned and applied the approach within 40 hours or a week’s work.

The student teams gained competence in two important areas. They were able to define good tasks for the test subjects, and they were able to express the problems they found in a clear and straightforward manner. Overall, this reflects competence in planning and writing. The students were less successful when it came to the identification of problems, which is the main purpose of a usability test.

REFERENCES


<table>
<thead>
<tr>
<th>Team No.</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of team members</td>
<td>6</td>
<td>7</td>
<td>7</td>
<td>6</td>
<td>7</td>
<td>7</td>
<td>8</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>6</td>
<td>8</td>
<td>7</td>
<td>5</td>
<td>5</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>No. of test subjects</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Tasks</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Test conduction</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>No. of problems</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Clarity of problem list</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Practical relevance of problem list</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>4</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Total number of identified problems</td>
<td>7</td>
<td>3</td>
<td>11</td>
<td>5</td>
<td>5</td>
<td>9</td>
<td>13</td>
<td>10</td>
<td>3</td>
<td>8</td>
<td>5</td>
<td>4</td>
<td>11</td>
<td>10</td>
<td>0</td>
<td>11</td>
<td>13</td>
<td>10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Team No.</th>
<th>19</th>
<th>20</th>
<th>21</th>
<th>22</th>
<th>23</th>
<th>24</th>
<th>25</th>
<th>26</th>
<th>27</th>
<th>28</th>
<th>29</th>
<th>30</th>
<th>31</th>
<th>32</th>
<th>33</th>
<th>34</th>
<th>35</th>
<th>36</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of team members</td>
<td>7</td>
<td>7</td>
<td>5</td>
<td>7</td>
<td>6</td>
<td>7</td>
<td>6</td>
<td>7</td>
<td>5</td>
<td>7</td>
<td>7</td>
<td>6</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>No. of test subjects</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Tasks</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>5</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Test conduction</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>No. of problems</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Clarity of problem list</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Practical relevance of problem list</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>5</td>
<td>2</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Total number of identified problems</td>
<td>12</td>
<td>11</td>
<td>7</td>
<td>14</td>
<td>16</td>
<td>5</td>
<td>11</td>
<td>6</td>
<td>6</td>
<td>8</td>
<td>11</td>
<td>5</td>
<td>1</td>
<td>8</td>
<td>5</td>
<td>5</td>
<td>12</td>
<td>12</td>
</tr>
</tbody>
</table>

Figure 7: All results of the empirical study with the 36 student teams. The first two rows in the table illustrate the total number of participants in each team and the number of used test subjects. Below five factors are listed with respect to the conduction of the entire test. The five factors are marked for each team with a value between 1 and 5 (where 5 is the best). At the bottom of the table the actual number of identified usability problems is listed for each team. The listing of the 36 teams is arbitrarily ordered.